BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Volumul 72 (76), Numărul 3, 2026
Secţia
CONSTRUCŢII DE MAŞINI
DOI: 10.2478/bipcm-2026-0028

RELIABILITY-SCREENED HUMAN EVALUATION OF FOUR CONVERSATIONAL AI PLATFORMS: CHATGPT, PERPLEXITY, CLAUDE AND GEMINI

BY

EMANUEL MIHĂLUȚE and NICOLAE-RĂZVAN MITITELU*

“Gheorghe Asachi” Technical University of Iași,
Department of Machine Manufacturing Technology, Iași, Romania

Received: August 3, 2026
Accepted for publication: August 20, 2026

Abstract. Human evaluation of large language models is widely used but its reliability is rarely verified. This study evaluated four conversational platforms, ChatGPT 5, Perplexity, Claude Sonnet 4 and Gemini 2.5 Flash, accessed on 20 August 2025. Twenty prompts covering factual, creative, technical and ethical-critical tasks were submitted identically to all four systems. Ten independent evaluators ranked each set of four responses on four criteria, yielding 800 forced-ranking blocks. Concordance was screened before any comparison. Only the factual set reached agreement above chance (Kendall's W = 0.0379; mean pairwise agreement 0.561); the creative, technical and ethical-critical sets were statistically indistinguishable from random ranking. On the factual set Perplexity and ChatGPT outranked Claude and Gemini (Friedman chi-square = 22.30, $p < 0.001$), an effect carried almost entirely by response structure. Three of four dimensions in routine use were therefore unusable as measurements. Reliability screening should precede interpretation in human evaluation of language models.

*Corresponding author; *e-mail*: nicolae-razvan.mititelu@academic.tuiasi.ro
© 2026 Emanuel Mihăluțe and Nicolae-Răzvan Mititelu
This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

Keywords: large language models; human evaluation; inter-rater reliability; forced ranking; Kendall's coefficient of concordance.

1. Introduction

Large language models (LLMs) represent one of the most consequential developments in applied computing of the last decade. The Transformer architecture introduced by (Vaswani *et al.*, 2017) removed the sequential bottleneck of recurrent networks and made large-scale pre-training tractable. Subsequently, the development of GPT models by OpenAI marked a defining milestone: GPT-2 (Radford *et al.*, 2019) and GPT-3 (Brown *et al.*, 2020) demonstrated for the first time the ability to generate coherent, long-length texts with a high level of stylistic and semantic consistency.

The evolution of these models was accelerated by the increase in data volume and available computing power, which enabled the emergence of new generation models such as GPT-4, Claude, Gemini, or systems based on the aggregation and processing of multiple sources of information, such as Perplexity AI. These tools have quickly become part of the global digital ecosystem, being integrated into productivity platforms, educational tools, and research applications.

The impact of LLMs is visible in multiple fields: in education, they are used to generate teaching materials and support the learning process (Kasneci *et al.*, 2023); in science, they facilitate the synthesis of academic literature; in industry, they are integrated into automated assistance, translation, and data analysis services (Dwivedi *et al.*, 2023). This diversification of uses has reinforced the perception that LLMs are not just technical tools, but emerging cognitive actors with major social and ethical implications.

Evaluating language models has been a methodological challenge from the outset. Among the most widely used automated benchmarks are MMLU (Hendrycks *et al.*, 2021), which contains over 15,000 questions from 57 academic domains, and BIG-Bench (Srivastava *et al.*, 2022), an extensive set of tasks that test reasoning, language and creativity. HELM (Liang *et al.*, 2022) first proposed a multidimensional evaluation, including not only accuracy but also robustness, transparency and fairness, drawing attention to the fact that a single metric is not sufficient to characterise the capabilities of an LLM (Mititelu *et al.,* 2025). The Chatbot Arena project introduced an interactive platform where users compare models in a head-to-head framework based on human votes.

Automated benchmarks cannot reach every property of interest. Creativity, explanatory depth and impartiality resist automatic scoring, so human judgement remains necessary (Mirowski *et al.*, 2023). Human evaluation, however, carries its own failure modes, and these are documented.

Howcroft *et al.* (2020) surveyed 165 natural language generation papers and found more than 200 distinct terms used for evaluated quality aspects, most of them undefined. Van der Lee *et al.* (2021) reported that a large share of human evaluations omit inter-annotator agreement entirely. Amidei *et al.* (2018) showed that agreement statistics in such tasks are frequently interpreted without reference to their chance baseline. An evaluation that never reports concordance cannot distinguish a real difference between systems from noise.

Two further constraints shape any comparison of commercial platforms. First, the endpoints are not static. Chen *et al.* (2024) measured substantial behavioural drift in GPT-3.5 and GPT-4 over three months on identical tasks, which makes an undated evaluation unreproducible in principle. Second, human and automated judges reward surface form. Zheng *et al*. (2023) documented position, verbosity and self-enhancement biases in pairwise judging, and Wu and Aji (2023) found that answers containing factual errors were rated more favourably than answers that were merely short or ungrammatical, across crowd-sourced annotators, expert annotators and model judges alike. Presentation and correctness are therefore not separable by default, and a rubric that does not force the separation will not achieve it.

A third constraint is architectural. Perplexity is a retrieval-augmented search product rather than a bare conversational model. Retrieval augmentation measurably changes factual behaviour: Lewis *et al*. (2020) showed that retrieval-augmented generation produces more factual language than a parametric-only baseline, and Shuster *et al*. (2021) reported over 60 percent fewer hallucinated responses under retrieval augmentation. Liu *et al*. (2023) audited four generative search engines, Perplexity among them, and found only 51.5 percent of generated sentences fully supported by their own citations, warning of a facade of trustworthiness. Comparing such a system against non-augmented assistants on factual questions is therefore not a like-for-like contrast, and the comparison must be reported as such.

The dimensions themselves are inherited from the evaluation literature rather than invented here. Factual accuracy, creativity, technical clarity and impartiality recur across benchmark and human-evaluation studies alike, and the ethical dimension in particular carries an established motivation: Bender *et al.* (2021) set out the risks of scale in language models, and Weidinger *et al.* (2021) organised the ethical and social harms into a systematic taxonomy. Detecting machine-generated text is itself unreliable with current tools (Mihăluţe and Mititelu, 2025), which places further weight on direct human assessment of model output.

Comparative studies of these same four platforms already exist. Jain (2024) reported a multifaceted assessment of ChatGPT, Gemini, Perplexity and Claude, and Chatbot Arena (Zheng *et al.*, 2023) supports continuous head-to-head comparison at far larger scale. The present study does not claim priority in comparing these systems. What it adds is the screening step: a stated

concordance threshold, computed against an exact chance baseline, applied before any difference between platforms is interpreted. Neither preference voting nor single-rater designs report that quantity, and the results below show what is lost when it is omitted.

This study addresses these constraints directly. It reports a structured, forced-ranking human evaluation of four platforms across four task families, and it makes inter-rater concordance a screening gate rather than an afterthought: a prompt set is interpreted only if its evaluators agreed above chance. The design contribution is the screening procedure. The empirical contribution is what that procedure reveals about which evaluation dimensions academic raters can and cannot apply.

2. Materials and Method

2.1. Evaluated systems and collection protocol

Four conversational platforms were evaluated. Model identifiers, versions and access conditions are given in Table 1. All prompts were submitted through each vendor public web interface on a single day, 20 August 2025, to remove version drift within the experiment. Default settings were used throughout. No system prompt, temperature or sampling parameter was modified. Each prompt was entered manually and identically on all four platforms. Responses were copied in full, unedited, into a single reference document, which is archived with the dataset.

Table 1

Evaluated systems, versions and access conditions

Platform	Model version	Vendor tier	Retrieval	Access date
ChatGPT	GPT-5	flagship	not enabled	20 Aug. 2025
Perplexity	not recorded	search product	enabled by design	20 Aug. 2025
Claude	Sonnet 4	mid tier	not enabled	20 Aug. 2025
Gemini	2.5 Flash	latency tier	not enabled	20 Aug. 2025

Note. The four systems do not occupy equivalent vendor tiers. Gemini 2.5 Flash is a latency-optimised tier rather than the Pro tier, and Claude Sonnet 4 is a mid-tier rather than Opus. Perplexity operates with web retrieval as an inherent property of the product. These asymmetries are treated as limitations in Section 4 and are not corrected for.

2.2. Prompt set and scoring criteria

Twenty prompts were used, in four sets of five: factual, creative, technical and ethical-critical. Each set was scored on four criteria specific to that set. The criteria differ between sets because the constructs differ. They are listed in full in Table 2, which the original protocol did not specify.

Table 2
Prompt sets and the four scoring criteria applied within each set

Prompt set	Prompts	Criterion 1	Criterion 2	Criterion 3	Criterion 4
Factual	P1-P5	Length	Detail	Structure	Academic tone
Creative	P6-P10	Originality	Expressivity	Storytelling	Imagery
Technical	P11-P15	Clarity	Correctness	Reasoning	Technical depth
Ethical-crit.	P16-P20	Impartiality	Critical depth	Structuring	Argument.

Note. The four criteria of the factual set describe properties of response form. None of them measures verified factual correctness against a ground truth. This is a construct limitation of the present design and it constrains the interpretation given in Sections 3.2 and 4.

2.3. Evaluator panel and forced-ranking protocol

Ten independent evaluators, E1 to E10, were recruited from academic staff and researchers with experience in academic discourse analysis and in the use of AI tools. Evaluators received written instructions specifying the criteria and the scoring rule. Participation was voluntary and unpaid, and evaluators were informed that their scores would be published in anonymised, aggregate form.

The scoring rule was a forced ranking. For each combination of prompt and criterion, an evaluator distributed the values 1, 2, 3 and 4 across the four platforms, where 1 denotes the weakest and 4 the strongest relative performance. Ties were not permitted. Forced ranking removes scale-use bias, because no evaluator can compress or inflate the whole scale, but it also makes the four platform scores constant-sum within every block. That property governs the analysis in Section 2.4.

The unit of observation is therefore the block, defined as one prompt by one criterion by one evaluator. Each prompt set contains 5 x 4 x 10 = 200 blocks and 800 individual scores. The whole experiment comprises 800 blocks and 3,200 scores.

2.4. Statistical analysis

Analysis proceeded in three stages, in a fixed order.

Stage 1, validity. Every block was checked against the forced-ranking rule. Six blocks out of 800 were not permutations of the set {1, 2, 3, 4} and were recorded as data-entry errors. No value was imputed. These blocks were excluded, leaving 196 valid blocks in the factual set, 198 in the creative set and 200 in each of the technical and ethical-critical sets.

Stage 2, reliability screening. Concordance among the ten evaluators was quantified for each prompt set with the Kendall coefficient of concordance W (Kendall and Babington Smith, 1939), the mean exact agreement over all 45 evaluator pairs, and the mean leave-one-out Spearman correlation between each evaluator and the consensus of the other nine. Because a block is a random permutation of four ranks under the null hypothesis, the chance baselines are known exactly: pairwise agreement is 0.250 and, for N blocks, the expected value of W is 1/N. The empirical distribution of the 24 possible permutations was additionally tested against uniformity by a chi-square goodness-of-fit test with 23 degrees of freedom. A prompt set was carried forward to Stage 3 only if it exceeded these baselines.

Stage 3, comparison. Forced ranking yields ordinal, constant-sum observations under a fully crossed repeated-measures design, so a between-groups analysis of variance is not admissible: the four platform scores are negatively dependent by construction and the observations are not independent. Differences between platforms were therefore tested with the Friedman rank test (Friedman, 1937), blocked on prompt, criterion and evaluator, followed by pairwise Wilcoxon signed-rank tests with Holm correction for the six comparisons. The significance threshold was 0.05. All computations were performed in Python 3 with SciPy and are reproducible from the deposited workbooks.

3. Results

3.1. Reliability screening

The four prompt sets separate sharply, and they separate before any platform comparison is made. Table 3 and Fig. 1 report the screening statistics.

Table 3

Inter-rater reliability by prompt set, against exact chance baselines

Prompt set	Blocks	W	Agreement	Rater rho	Unif. p	Pass
Factual	196	0.0379	0.561	+0.810	< 0.001	yes
Creative	198	0.0086	0.246	-0.127	0.660	no
Technical	200	0.0025	0.248	-0.052	0.811	no
Ethical-crit.	200	0.0031	0.250	-0.057	0.799	no
Chance baseline	-	0.005	0.250	0.000	> 0.05	-

Note. Blocks = valid blocks. W = Kendall coefficient of concordance. Agreement = mean exact agreement over the 45 evaluator pairs. Rater rho is the leave-one-out Spearman correlation between an evaluator and the consensus of the other nine, averaged over evaluators. Uniformity p is the chi-square goodness-of-fit of the observed rank permutations against a uniform distribution over all 24 permutations; a high value means the data cannot be distinguished from random ranking. Pass indicates whether the set was carried forward to Stage 3.

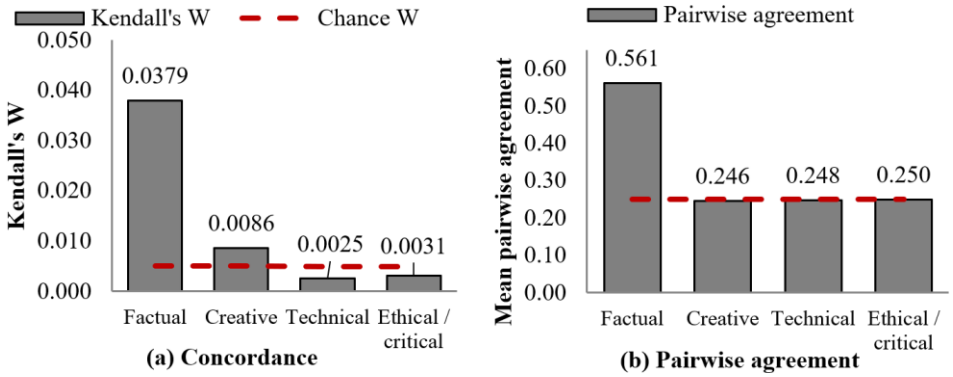


Fig. 1 – Concordance (a) and mean pairwise agreement (b) by prompt set, against chance.

In the factual set every one of the ten evaluators correlates positively with the consensus of the other nine, with individual coefficients between +0.71 and +0.98. Mean pairwise exact agreement is 0.561, more than double the chance value of 0.250, and the permutation distribution departs strongly from uniformity.

In the creative, technical and ethical-critical sets no evaluator exceeds +0.10 against the consensus, pairwise agreement sits at 0.246, 0.248 and 0.250 against a chance value of 0.250, and the observed permutations are uniformly distributed across all 24 possibilities. Three artefactual explanations were tested on these sets and rejected. No evaluator was reverse-coded, since no leave-one-out correlation approaches the negative values that scale inversion would produce. Column misassignment was excluded, because allowing each evaluator platform labels to be permuted freely so as to maximise agreement raised W no higher than it rises on simulated random data. Neither fatigue nor a single defective criterion is responsible, because concordance is flat across the five prompts and across the four criteria within each set.

These three sets therefore carry no recoverable evaluator signal, and no comparison between platforms is reported from them beyond the descriptive record in Table 5.

3.2. Factual prompt set

On the one set that passed screening, the four platforms differ: Friedman chi-square(3) = 22.30, $p < 0.001$, with N = 196 blocks. Mean ranks were Perplexity 2.776, ChatGPT 2.648, Gemini 2.321 and Claude 2.255. Table 4 reports the six Holm-corrected pairwise comparisons.

Table 4

Factual set: pairwise comparisons, Wilcoxon signed-rank with Holm correction

Comparison	Rank diff.	p (raw)	p (Holm)	Signif.
Perplexity vs. Claude	+0.520	< 0.001	< 0.001	yes
Perplexity vs. Gemini	+0.454	0.0015	0.0076	yes
ChatGPT vs. Claude	+0.393	0.0082	0.0330	yes
ChatGPT vs. Gemini	+0.327	0.0264	0.0793	no
ChatGPT vs. Perplexity	-0.128	0.1555	0.3109	no
Claude vs. Gemini	-0.066	0.4644	0.4644	no

Perplexity outranks both Claude and Gemini, and ChatGPT outranks Claude. Perplexity and ChatGPT do not separate from each other, and neither do Claude and Gemini. The result is therefore a two-tier ordering rather than a strict ranking of four systems.

Decomposing by criterion locates the effect precisely, as shown in Fig. 2. Structure alone accounts for most of it: $W = 0.346$, chi-square(3) = 51.96, $p < 0.001$, with Perplexity at 3.32 and Claude at 1.74. Detail contributes weakly but significantly, $W = 0.080$, $p = 0.007$. Length, with $W = 0.017$ and $p = 0.489$, and academic tone, with $W = 0.030$ and $p = 0.231$, do not reach significance and their concordance is at chance. Evaluators agreed reliably on how a response was organised, marginally on how much it contained, and not at all on how long it was or what register it used.

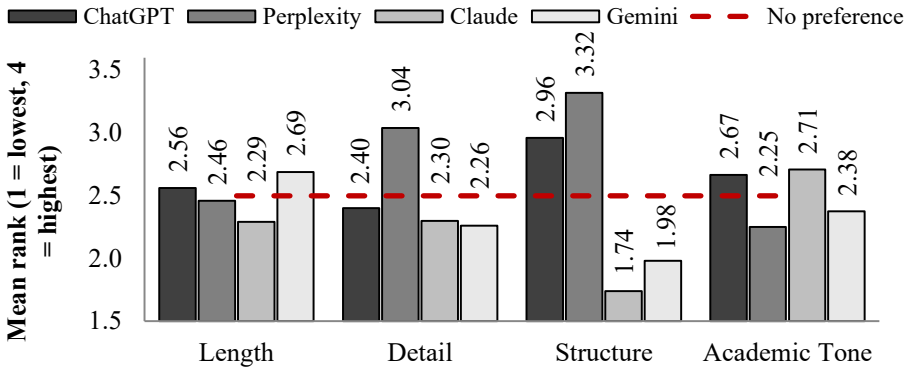


Fig. 2 – Factual set: mean ranks by criterion, with concordance for each criterion.

Two observations follow. First, length is an objectively countable quantity, yet evaluators did not converge on it. This bounds how carefully the responses were inspected even within the set that passed screening. Second,

none of the four criteria measures factual correctness. The finding is that Perplexity and ChatGPT produced better-organised answers, not that they produced more accurate ones.

3.3. Prompt sets that failed screening

Table 5 records the descriptive outcome of the three sets that did not pass Stage 2. None reaches significance, and none should be interpreted, because the underlying scores are indistinguishable from random assignment. The values are reported for completeness and to support reuse of the deposited data.

Table 5

Prompt sets that failed reliability screening: descriptive record only

Prompt set	ChatGPT	Perplex.	Claude	Gemini	chi-sq.(3)	p
Creative	2.480	2.535	2.348	2.636	5.13	0.162
Technical	2.485	2.470	2.450	2.595	1.52	0.678
Ethical-crit.	2.465	2.500	2.435	2.600	1.85	0.603

Note. Values are mean ranks on the 1 to 4 scale. No comparison in this table is interpretable, because these sets did not pass the reliability gate in Table 3.

4. Discussion

The principal result of this study is methodological. Of four evaluation dimensions in routine use, only one produced agreement above chance among ten academic evaluators working from written instructions. Had concordance not been screened, all four sets would have been reported as comparable measurements, and three of them would have carried conclusions built on noise. This is the failure mode that Howcroft *et al.* (2020) and Van der Lee *et al.* (2021) identified in the wider literature, observed here within a single controlled design.

The pattern of failure is informative rather than random. The dimension that survived, factual response form, is anchored in observable properties of the text. The dimensions that failed, creativity, technical quality and ethical reasoning, require the evaluator to supply a standard that the instructions did not provide. Asking ten people to rank four responses for impartiality on a scale from 1 to 4, with no anchor for what each point means, produces ten private scales that do not coincide. Amidei *et al.* (2018) argue that divergence among annotators is partly ineliminable. The present data suggest that a substantial part of it is instead a rubric artefact, and therefore remediable by anchoring.

On the factual set the ordering is consistent with the architectural difference between the systems. Perplexity retrieves and cites, and it emits

headed, source-marked text. Retrieval augmentation improves factual grounding (Lewis *et al.*, 2020; Shuster *et al.*, 2021), so a genuine advantage on factual tasks is plausible. The present data cannot confirm it, because the rubric scored form rather than correctness, and the effect concentrates almost entirely in the structure criterion. Wu and Aji (2023) showed that judges reward presentation over substance, and Liu *et al.* (2023) found that generative search engines fully support only about half of their own cited sentences. The conservative reading of the present result is that evaluators reliably detected the response format characteristic of a retrieval product. That same format is also what would identify the platform to an unblinded evaluator, so the two explanations cannot be separated in this design.

The comparison is further constrained by tier asymmetry. GPT-5 was a flagship-tier model at the access date, whereas Claude Sonnet 4 and Gemini 2.5 Flash sit below their vendor top tiers. Any ordering reported here describes the four configurations actually tested, not the four vendors best available systems.

4.1. Limitations

Five limitations bound the conclusions. First, evaluators were not blinded to platform identity and response order was not randomised, so both position bias and brand priors are uncontrolled (Zheng *et al.*, 2023). Second, the factual rubric measures form, not verified accuracy, because no answer key was used. Third, the model configurations are unequal in vendor tier, and Perplexity retrieval cannot be disabled. Fourth, the Perplexity underlying model was not recorded. Fifth, the results describe a single access date, and commercial endpoints drift, so the ordering is a dated snapshot and is not expected to replicate against current model versions.

4.2. Recommendations for future work

Three changes would address these limitations without enlarging the study. Each criterion should be given a written anchor for every scale point, and a calibration round should be run on a subset of prompts, with concordance verified before the main evaluation proceeds. Responses should be stripped of vendor-identifying markers, including citation footers and characteristic formatting, and presented in randomised order. For the factual dimension, a documented answer key should be prepared from the archived responses so that correctness is scored independently of form. Because the responses are archived verbatim, all three changes can be implemented on the existing corpus without access to the original model versions.

5. Conclusions

This study evaluated four conversational AI platforms with a forced-ranking protocol and ten independent evaluators, and screened inter-rater concordance before interpreting any comparison.

Three of the four evaluation dimensions failed the screen. Creative, technical and ethical-critical rankings were statistically indistinguishable from random assignment, with concordance at or below the chance baseline and rank permutations uniformly distributed. Reverse coding, column misassignment and rater fatigue were tested and excluded as explanations. The most probable cause is the absence of anchored definitions for the scale points.

The factual dimension passed, with concordance well above chance and every evaluator aligned with the consensus. On that dimension Perplexity and ChatGPT outranked Claude and Gemini, with a Friedman chi-square of 22.30 and $p < 0.001$. The effect is carried by the structure criterion and does not extend to verified correctness, which this design did not measure.

The practical implication is procedural. Inter-rater concordance should be computed and reported before differences between models are interpreted, and it should be compared against the exact chance baseline that the scoring scheme implies. A structured human evaluation that omits this step cannot distinguish a measured difference from an artefact of the rubric.

Data availability. The 20 prompts, the verbatim responses of all four platforms, and the complete 3,200-score matrix together with the reliability and Friedman analyses are deposited at https://doi.org/10.5281/zenodo.17209054. The six invalid blocks are flagged in the deposited workbooks and were excluded from all analyses.

REFERENCES

Amidei J., Piwek P., Willis A., *Rethinking the Agreement in Human Evaluation Tasks*, Proc. of the 27th Internat. Conf. on Computat. Linguistics, COLING 2018, August 2018, Santa Fe, USA, Association for Computational Linguistics, 3318-3329 (2018).

Bender E.M., Gebru T., McMillan-Major A., Mitchell M., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, Proc. of the 2021 ACM Conf. on Fairness, Accountability, and Transparency, FAccT 2021, March 2021, Virtual Event, Canada, ACM, New York, 610-623 (2021).

Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., *Language Models are Few-Shot Learners*, Adv. in Neural Inform. Process. Syst., **33**, 1877-1901 (2020).

Chen L., Zaharia M., Zou J., *How Is ChatGPT's Behavior Changing over Time?*, Harvard Data Sci. Rev., **6**, *2* (2024).

Dwivedi Y.K., Kshetri N., Hughes L., Janssen M., Drăgoicea M., Wright R., *So What if ChatGPT Wrote It? Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative AI for Research, Practice and Policy*, Internat. J. of Inform. Manag., **71**, 102642 (2023).

Friedman M., *The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance*, J. of the Amer. Statist. Assoc., **32**, *200*, 675-701 (1937).

Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., Steinhardt J., *Measuring Massive Multitask Language Understanding*, Proc. of the 9th Internat. Conf. on Learning Representations, ICLR 2021, May 2021, Virtual Event (2021).

Howcroft D.M., Belz A., Clinciu M.-A., Gkatzia D., Hasan S.A., Mahamood S., Mille S., van Miltenburg E., Santhanam S., Rieser V., *Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions*, Proc. of the 13th Internat. Conf. on Natural Language Generation, INLG 2020, December 2020, Dublin, Ireland, Association for Computational Linguistics, 169-182 (2020).

Jain S., *A Comparative Assessment of Advanced Conversational Agents: A Multifaceted Evaluation of ChatGPT, Gemini, Perplexity, and Claude*, Internat. J. of Emerg. Technol. and Adv. Engng., **14**, *2*, 59-69 (2024).

Kasneci E., Sessler K., Küchemann S., Bannert M., Dementieva D., Fischer F., Kasneci G., *ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education*, Learning and Individ. Differences, **103**, 102274 (2023).

Kendall M.G., Babington Smith B., *The Problem of m Rankings*, The Ann. of Math. Statist., **10**, *3*, 275-287 (1939).

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-t., Rocktäschel T., Riedel S., Kiela D., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, Adv. in Neural Inform. Process. Syst., **33**, 9459-9474 (2020).

Liang P., Bommasani R., Lee T., Tsipras D., Soylu D., Yasunaga M., Zhang Y., Narayanan D., Wu Y., Kumar A., *Holistic Evaluation of Language Models*, arXiv:2211.09110 (2022).

Liu N.F., Zhang T., Liang P., *Evaluating Verifiability in Generative Search Engines*, Findings of the Assoc. for Computat. Linguistics, EMNLP 2023, December 2023, Singapore, Association for Computational Linguistics, 7001-7025 (2023).

Mihăluţe E., Mititelu N.-R., *The Limitations of Existing Software in Identifying Content Created with the Help of AI*, Bul. Inst. Polit. Iaşi, **71**, *3*, s. Constr. de Maşini, 9-18 (2025).

Mirowski P., Mathewson K.W., Pittman J., Evans R., *Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals*, arXiv:2209.14958 (2023).

Mititelu N.-R., Hobjala R., Ermolai V., Rîpanu I.M., Bisog C., *Use of AI Language Models in Generating 3D Printable Models*, Acta Technica Napocensis - Series: Applied Mathematics, Mechanics, and Engineering, vol. 68, no. 1S. Available: https://atna mam.utcluj.ro/index.php/Acta/article/view/2779. (2025)

Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I., *Language Models are Unsupervised Multitask Learners*, OpenAI Technical Report (2019).

Shuster K., Poff S., Chen M., Kiela D., Weston J., *Retrieval Augmentation Reduces Hallucination in Conversation*, Findings of the Assoc. for Computat. Linguistics, EMNLP 2021, November 2021, Punta Cana, Dominican Republic, Association for Computational Linguistics, 3784-3803 (2021).

Srivastava A., Rastogi A., Rao A., Shoeb A.A.M., Abid A., Fisch A., Brown A.R., Santoro A., Gupta A., Garriga-Alonso A., *Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models*, arXiv:2206.04615 (2022).

Van der Lee C., Gatt A., van Miltenburg E., Krahmer E., *Human Evaluation of Automatically Generated Text: Current Trends and Best Practice Guidelines*, Comput. Speech and Lang., **67**, 101151 (2021).

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I., *Attention Is All You Need*, Adv. in Neural Inform. Process. Syst., **30**, 5998-6008 (2017).

Weidinger L., Mellor J., Rauh M., Griffin C., Uesato J., Huang P.-S., Cheng M., Glaese M., Balle B., Kasirzadeh A., *Ethical and Social Risks of Harm from Language Models*, arXiv:2112.04359 (2021).

Wu M., Aji A.F., *Style Over Substance: Evaluation Biases for Large Language Models*, arXiv:2307.03025 (2023).

Zheng L., Chiang W.-L., Sheng Y., Zhuang S., Wu Z., Zhuang Y., Lin Z., Li Z., Li D., Xing E.P., Zhang H., Gonzalez J.E., Stoica I., *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*, Adv. in Neural Inform. Process. Syst., **36**, Datasets and Benchmarks Track (2023).

EVALUARE UMANĂ CU VERIFICAREA FIDELITĂȚII A PATRU PLATFORME CONVERSAȚIONALE DE INTELIGENȚĂ ARTIFICIALĂ: CHATGPT, PERPLEXITY, CLAUDE ȘI GEMINI

(Rezumat)

Evaluarea umană a modelelor lingvistice mari este larg utilizată, însă fidelitatea ei este rareori verificată. Lucrarea evaluează patru platforme conversaționale, ChatGPT 5, Perplexity, Claude Sonnet 4 şi Gemini 2.5 Flash, accesate la 20 august 2025. Douăzeci de solicitări din domeniile factual, creativ, tehnic şi etic-critic au fost transmise identic celor patru sisteme. Zece evaluatori independenți au ierarhizat fiecare set de patru răspunsuri după patru criterii, rezultând 800 de blocuri de ierarhizare forțată. Concordanța a fost verificată înaintea oricărei comparații. Numai setul factual a depășit pragul întâmplării (coeficientul W al lui Kendall = 0,0379; acord mediu între perechi 0,561); seturile creativ, tehnic şi etic-critic au fost statistic imposibil de deosebit de o ierarhizare aleatorie. Pe setul factual, Perplexity şi ChatGPT au depășit Claude şi Gemini (testul Friedman, chi-pătrat = 22,30, $p < 0,001$), diferența fiind determinată aproape integral de structura răspunsului. Verificarea fidelității trebuie să precedă interpretarea în evaluarea umană a modelelor lingvistice.